Exhibit 12

MACHINE LEARNING, USER EXPERIENCE

May 11, 2022

On-device voice control on Sonos speakers

Alice Coucke

Head of Machine Learning Research, Voice Experience

Joseph Dureau

Vice President, Voice Experience

David Leroy

Director, Voice Experience Machine Learning

Sébastien Maury

Senior Director, Voice Experience

Context

Sonos Voice Control

Sonos is committed to delivering new experiences that effortlessly connect listeners to the content they love. One of the most natural ways to connect to your music is with your voice, yet many customers who have purchased our voice-capable speakers are choosing not to activate or use voice services. We hear from our customers that privacy concerns and failure to meet expectations for accuracy, speed, and ease of use are common reasons for this. As of June 1st, Sonos Voice Control will deliver the experience our customers want without compromise - one that addresses speed, accuracy and privacy equally.

Sonos Voice Control is a voice interface for Sonos users to control their music and their Sonos system, placing privacy at the heart of its design. It works just like the app: no additional data leaves the home. It grants users the ability to initiate music or radio, control the volume, navigate the tracks, target or group Sonos speakers, manage their playlists and more. Sonos Voice Control can run on any microphone-enabled Sonos speaker, including older generations, in both WiFi and Bluetooth modes.

Our voice experience team has been working on multiple aspects of speech understanding: speech enhancement, wake word detection, automatic speech recognition, and natural language understanding. The core development has been in Paris, within Sonos' first product and engineering site in Europe. To bring this feature to life, the team brings together an interdisciplinary group of software engineers (embedded, backend, frontend, quality assurance) and linguists in addition to signal processing and machine learning scientists.

In this post, we'll tell you more about voice assistants: how they usually work and their current limitations. We'll present the approach we have chosen at Sonos, focused on a fast and accurate vocal gateway to your listening experience, without compromising your privacy. And of course, we'll also share with you some of the main challenges we faced along the way.

Smart speakers and their current limitations

Anatomy of common voice assistants

Typical voice assistants map an input voice request captured by the microphone on the speaker to an action performed on the speaker. This overall task is achieved by combining and chaining several components.

First, the **audio front-end** processes the signals from the speaker's microphones and applies a series of transformations in order to get the final, enhanced audio signal passed on to the downstream components. The goal of the audio front-end is to clean the input signal and

remove parasitic interferences, such as background noise, reverberation due to the reflection of the sound on the walls or objects placed within the room, and especially self-sound (the music playback from the speaker itself). Such speech enhancement systems typically

music playback from the speaker itself). Such speech enhancement systems typically combine digital signal processing techniques incorporating domain knowledge about physics and perception, together with machine learning models to increase the model's power to generalize. The role of the audio front-end is illustrated on the following two audio files:

0:00 / 0:07

0:00 / 0:07

Next, a **wake word detector** is continuously listening for a predefined keyword in the cleaned audio stream , to initiate an interaction with the voice assistant. The wake word detector runs on the speaker itself (sometimes with additional verification steps in the cloud), and it is usually a deep neural network that maps an audio sequence to a binary output: whether or not the wake word is present inside the audio input. The wake word, like the audio front-end, processes every chunk of voice captured by the microphone. As soon as the wake word is detected, the most common behavior is to stream the subsequent audio directly to a remote server where the next steps are performed.

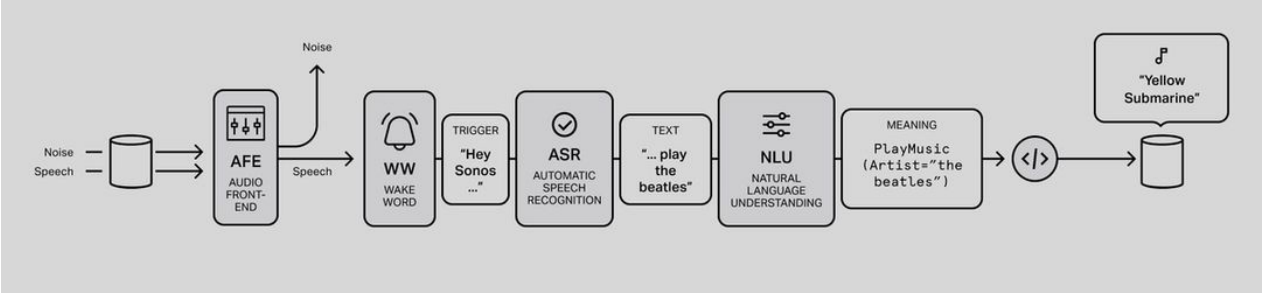
Then, an **Automatic Speech Recognition** (ASR) model transcribes the subsequent speech query into text. It predicts the sequence of words from a sequence of audio frames. On one hand, traditional approaches to ASR consist of decomposing this task into an acoustic part, modeling the probability of an audio sequence given a word sequence, and a language part modeling the probability of a word sequence. This decoupled method enables training the acoustic and the language models separately: the former on generic speech audio data with topic-independent text transcripts, and the latter on text sentences in the voice assistant's scope (e.g. music control, weather, timers). A decoupled ASR is therefore easier to fine tune.

The language model is the largest part of these traditional ASR systems; its size and computational footprint increase exponentially with the voice assistant's scope. Large-vocabulary language models for general purpose voice assistants can be up to terabytes in size. On the other hand, the speech recognition field is seeing an increase in the use of end-to-end models, which directly represent the probability of a word sequence given an audio sequence, performing acoustic and language modeling together. These models are more compact but much harder to customize and adapt to rare pronunciations of artists or song names from different languages, for instance.

Finally, a **Natural Language Understanding** module extracts information from the decoded text in order to interpret the user's request. This information consists of two major items: the

overall intent (e.g. to play a song or adjust the volume) and the slots (e.g. the artist or song name, the volume level). These items may be extracted jointly with one system, or sequentially, in which case we refer to the tasks as intent classification and slot filling.

Subsequent modules are responsible for performing the requested action itself, that may require searching for content on a cloud-based service, or calling a text-to-speech (or speech synthesis) component that generates an audio response from the speaker in order to provide information, elicit a response, or initiate a dialogue.



Typical voice assistant mapping a customer's voice request to an action performed on the speaker.

Typical voice assistants still mostly live in the cloud and use the wake word as a gateway to powerful remote servers, removing pressure on the memory and computational footprint of the underlying machine learning models. However, accuracy and privacy are growing concerns among connected devices users, as discussed in the following.

Concerns around voice assistants

One of the most natural ways our customers can connect to their content is through voice, yet many of Sonos' voice-capable speakers aren't being used for that purpose today. Indeed, we hear from our customers that privacy concerns and disappointment with the global experience mean that they are not activating voice on compatible Sonos products (Sonos beta survey, 2019). Moreover, a recent study shows that smart speaker adoption is plateauing in the US (Futuresource Virtual Assistant Tracker, 2021). The reason is twofold.

Far-field voice recognition is still a challenging problem

First, voice recognition performance is strongly degraded in acoustic conditions other than what is often referred to as "clean close field", meaning the user is speaking 30 cm or less from the microphone, without background noise or competing speech. Dictation on smartphones or in-car voice assistants are generally easier use cases than voice interactions with a smart speaker. On a smart speaker, the signal's quality may be degraded in various ways: attenuation during the propagation of sound between the source and the speaker's microphone (to make things worse, the location of the speaker can't be assumed to be fixed

microphone (to make things worse, the location of the speaker can't be assumed to be fixed over time, in particular for battery-powered portable speakers like Roam or Move), non-stationary background noise interferences, reverberation on walls and furniture, and self-sound [1]. Characteristics like the room geometry, the composition of the furniture, position of the user with respect to the speaker, etc. are impossible to predict and the number of combinations is almost infinite in the general case for a room in a customer's house. On the contrary, the geometry of a given car is usually known in advance, the background noise more predictable, and the microphone is located at a fixed position.

A series of scientific challenges recently aimed at better defining and gathering the efforts of the scientific community on far-field voice recognition, namely the REVERB challenge [2], the series of CHiME challenges [3]–[6], and the ASpIRE challenge [7]. Solving this task is a requisite for any industrial application of voice assistants, but requires a lot of resources (data, compute, and expertise). Currently, only the biggest tech companies have been able to develop far-field, general purpose voice assistants.

Privacy is a barrier for smart speaker adoption

Second, privacy is often listed as the main barrier to voice adoption, and this trend is increasing. It is the number one reason for not acquiring a smart speaker (Qualcomm State of Play report, 2019) or activating a voice assistant on Sonos speakers (Sonos beta survey, 2019). Moreover, 43% of US audio product owners are put off by smart speakers for privacy reasons (Futuresource Audio Tech Lifestyles, 2020) and always-on listening is an issue for a growing proportion of smart speakers' detractors (36% in 2017 to 55% in 2019, Smart Audio Report 2020).

Privacy issues with smart speakers stem from different sources. First and foremost, unintended voice assistant triggers are unavoidable [8, 9] (see these two studies from 2020 on accidental voice assistant triggers), leading to potentially private information being transferred without the active control or consent of the user. The always-on microphone should in principle only detect a predefined wake word before streaming audio directly to the cloud, but false alarms happen. Every smart speaker user has experienced a voice assistant talking without being summoned or responding to an incorrectly interpreted query. In many other cases, the wake word triggers silently: the customer is not aware that audio from their home is currently being streamed, processed, or potentially stored on a remote server. It is impossible to guarantee zero false alarms due to the statistical nature of wake word detectors that are based on deep neural networks. These models build latent representations in high dimensional spaces, that cause them to sometimes learn spurious correlations and trigger on specific audio features not perceptible by a human ear, or sounds that have nothing to do with the wake word [8].

Most voice assistants require large scale data collection to deliver good performance. The stored audio clips of voice interactions from customers that opt into data collection are

transcribed by human annotators to determine with confidence what they contain. Most of the artificial intelligence in the industry is still highly *supervised*, which means manually-provided ground truth labels are needed to train the underlying machine learning algorithms and improve them.

These issues are amplified because voice is a biometric marker: it is a unique identifier that can be used to identify or impersonate someone, and as opposed to a password it is something that you cannot change. Moreover, scientists have shown that someone's identity [10], intention [11], emotional state and pathological condition [12] may be retrieved to a great extent from audio clips of their voice. Audio and speech anonymization is still an open scientific question and an active field of research [13]-[16].

To alleviate these privacy concerns, efforts are being made to move the decoding away from the cloud and closer to the user's personal connected device. So far, they have been mostly concentrated on smartphones, which are much more powerful than a typical smart speaker in terms of computation.

Introducing Sonos Voice Control

We are offering an alternative approach to voice control, addressing each key area of concern. The unbounded value proposition of general purpose voice assistants makes it a potentially intractable problem that may justify large scale data collection to get a satisfactory user experience. We are taking a different, more focused approach, that doesn't require transmitting or storing any audio data from any customer.

Our approach

Sonos Voice Control focuses on music control and on the listening experience for the Sonos ecosystem. Music is the top use case for voice-enabled smart speakers (Futuresource Audio Tech Lifestyles, 2021). Addressing a bounded use case, rather than general purpose interactions, means that there is no need for a "discovery" type of feedback loop in which user expectations and desires are scrutinized through the transcripts of unsupported queries. It also allows Sonos Voice Control to get high in-domain accuracy, especially in challenging acoustic environments. Focusing on music alleviates the need for large scale data collection, to the extent that training and evaluation data may be collected from third-party providers or agencies that hire people who are willingly and knowingly reading scripted queries, or voluntary data from sources like Sonos Beta. To keep up-to-date with the customer demand, this training and evaluation data only needs to be scaled with new entries integrating the

music catalog that evolves with time. The phrasings and formulations used to request music on the other hand are limited in variety, and are not expected to vary significantly over time. Moreover, our approach to ASR that decouples the acoustic modeling from the language modeling means that we mainly need to scale coverage of music entities on the text side, which is much easier than scaling audio recordings.

The voice recognition stack runs directly on Sonos speakers. The voice of the user is processed locally and is never sent to a centralized cloud server. The machine learning models are trained on scripted audio data and then deployed for inference on the smart speakers. Potential wake word false triggers are therefore not a privacy concern anymore, because no audio data is sent to the cloud after the wake word is detected. This embedded approach is arguably the most straightforward and transparent way to perform privacy-preserving speech recognition. It is simpler to trust and comprehend for non-experts than other approaches like differential privacy, or speech anonymization, where a statistically positive amount of information still leaves people's homes to feed and improve centralized machine learning models.

This focused and embedded approach comes with many benefits.

Privacy: Closed domain models are smaller and require less training data. With Sonos Voice Control, we show that large scale personal data collection is not a requirement to provide a satisfying user experience on a single yet very challenging domain. The Sonos voice engine processes your voice and understands your requests entirely on the speaker. No audio or transcript is sent to the cloud, stored, listened to, read or labeled by anyone, so all the conversations in a user's home remain in the home.

Follow-up requests: This approach enables the Sonos voice engine to keep locally processing the audio stream even after the users are done with their initial voice request, without compromising their privacy. As a result, simple additional commands (like adjusting the volume or skipping a song) may be understood without repeating the wake word. This functionality, when present, is only an opt-in on other voice assistants.

Personalization: The embedded approach makes it natural for every device to get their own voice engine instead of having a single model on a centralized cloud server. Therefore, the local model can be adapted to the content of the customer's personal library. That means it will understand the names of the songs and artists they've saved and liked or the playlists they've created on their preferred music streaming services.

Bringing Sonos Voice Control to life raised a lot of technical challenges that we'll address to conclude this post. Our team already had significant expertise in this area, having joined Sonos from Snips, a French start-up that specialized in private-by-design voice interfaces. The team authored several scientific publications [17]-[22] in speech recognition and made publicly available several audio datasets for research purposes, today widely used by the

speech community (on spoken language understanding, wake word detection, and open-vocabulary keyword spotting).

The challenges we faced

Footprint: One of the key challenges to perform speech recognition directly on Sonos speakers is to be able to run all the operations needed to process user voice requests in real-time, using the computing and memory resources dedicated to Sonos Voice Control. We are very proud to have been able to deploy it on every model of voice-enabled Sonos speakers ever sold, including the most limited Sonos One Gen 1 with only a fraction of a Cortex-A9 CPU with 1GB of RAM available. It required handcrafted software optimizations for each target speaker to make sure we maximize the compute resource usage. This challenge has been achieved by optimizing a trade-off between accuracy and computational efficiency when designing the acoustic model, and by contextualizing the language model and the natural language understanding component to the music domain, in order both to reduce their size and increase their in-domain accuracy. We also developed our own deep neural network inference library to better control the inference process, especially matrix multiplication operations, to improve real time performance (see this series of posts on the topic on Sonos Tech Blog). We managed to keep the latency of Sonos Voice Control on par with cloud-based voice assistants, despite significant constraints on the hardware environment.

Music catalog management: While the phrasings and formulations used to request music are limited in variety, this use case is still very complex and involves an immense vocabulary of constantly evolving multilingual music entities. To make sure that common content is correctly decoded, regardless of, e.g., the language of origin of a given song name, each music entity is processed through a curation procedure. This automated content processing and annotation pipeline ensures that relevant spelling and pronunciations are generated, so that the content can be recognized even when it is pronounced in different ways by users and that we are robust to deviations from regular pronunciations that cannot be listed or known beforehand. On top of a common catalog of the most popular music entities, we also locally include each user's personal library in their own voice engine. This personal content is favored over any other content in the resolution strategy in order to maximize the content coverage for everyone. The resulting music catalog is updated regularly.

Acoustic and demographic robustness: The machine learning models are tuned to adapt to each Sonos speaker's acoustic properties, from portable all-in-one players to home-theater soundbars, and to the acoustic environments typical of a Sonos user's home: the command being given several meters away from the speaker, with external background noise (e.g. conversations or TV) or music played by the speaker itself (self-sound). To do so without large scale personal data collection, we rely heavily on scripted audio collection and simulations of large numbers of acoustic environments to artificially generate data representative of the

production environment. These simulations are based on randomly generated virtual acoustic rooms with random speaker locations, sound pressure level adjustment, and addition of external noise and self-sound with controlled intensity [23]. Voluntary data from the Sonos Beta community helps us make sure that these simulations are representative of the production environment and improve our models. Finally, to ensure that Sonos Voice Control works well for everyone, the performance of our voice engine has been optimized for the main US dialectal regions for both native and non-native speakers on a very large dataset balanced in terms of age, gender, and accent.

Conclusion

With Sonos Voice Control, we want to provide a new experience that delivers fast, accurate, and hands-free control of your music and your Sonos system with unmatched privacy. This is an exciting adventure that started years ago and we are really eager to hear feedback from our customers to learn, iterate, and continue improving the product. We'll keep sharing on Sonos Tech Blog about the encountered challenges and our proposed solutions in the future, so stay tuned!

References

[1] Haeb-Umbach, Reinhold, et al. "Far-field automatic speech recognition." Proceedings of the IEEE 109.2 (2020): 124-148.
[2] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. HaebUmbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," EURASIP Journal on Advances in Signal Processing, 2016.
[3] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third "CHiME" speech separation and recognition challenge: Analysis and outcomes," Computer Speech and Language, vol. 46, pp. 605–626, Nov. 2017.
[4] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," Computer Speech and Language, 2016.
[5] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in Proc. of Annual Conference of the International Speech Communication Association (Interspeech), 2018.
[6] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," CoRR, 2020.
[7] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Dec 2015, pp. 547–554.
[8] Dubois, Daniel J., et al. "When speakers are all ears: Characterizing misactivations of iot smart speakers." Proceedings on Privacy Enhancing Technologies 2020.4 (2020).
[9] Schönherr, Lea, et al. "Unacceptable, where is my privacy? exploring accidental triggers of smart speakers." arXiv preprint arXiv:2008.00508 (2020).
[10] Reynolds, Douglas A. "Speaker identification and verification using Gaussian mixture speaker models." Speech communication 17.1-2 (1995): 91-108.

mixture speaker models." Speech communication 17.1-2 (1995): 91-108.
[11] Gu, Yue, et al. "Speech intention classification with multimodal deep learning." Canadian conference on artificial intelligence. Springer, Cham, 2017.
[12] Gómez-Vilda, Pedro, et al. "Glottal source biometrical signature for voice pathology detection." Speech Communication 51.9 (2009): 759-781.
[13] Hashimoto, Kei, Junichi Yamagishi, and Isao Echizen. "Privacy-preserving sound to degrade automatic speaker verification performance." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.
[14] Qian, Jianwei, et al. "Voicemask: Anonymize and sanitize voice input on mobile devices." arXiv preprint arXiv:1711.11460 (2017).
[15] Jin, Qin, et al. "Speaker de-identification via voice transformation." 2009 IEEE Workshop on Automatic Speech Recognition & Understanding. IEEE, 2009.
[16] Srivastava, Brij Mohan Lal, et al. "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?." arXiv preprint arXiv:1911.04913 (2019).
[17] Coucke, Alice, et al. "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces." 2018 First Workshop on Privacy in Machine Learning and Artificial Intelligence (PiMLAI'18). FAIM Workshop, 2018.
[18] Coucke, Alice, et al. "Efficient keyword spotting using dilated convolutions and gating." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
[19] Leroy, David, et al. "Federated learning for keyword spotting." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
[20] Saade, Alaa, et al. "Spoken language understanding on the edge." 2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS). IEEE, 2019.

[21] D'Ascoli, Stéphane, et al. "Conditioned Text Generation with Transfer for Closed-Domain Dialogue Systems." International Conference on Statistical Language and Speech Processing. Springer, Cham, 2020.
[22] Bluche, Théodore, and Thibault Gisselbrecht. "Predicting Detection Filters for Small Footprint Open-Vocabulary Keyword Spotting." Interspeech 2020. ISCA-International Speech Communication Association, 2020.
[23] Bezzam, Eric, et al. "A study on more realistic room simulation for far-field keyword spotting." 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2020.

Continue reading in Machine Learning:

MACHINE LEARNING, AUDIO PROCESSING

Sonos at ICASSP 2022

Read More →

July 27, 2022

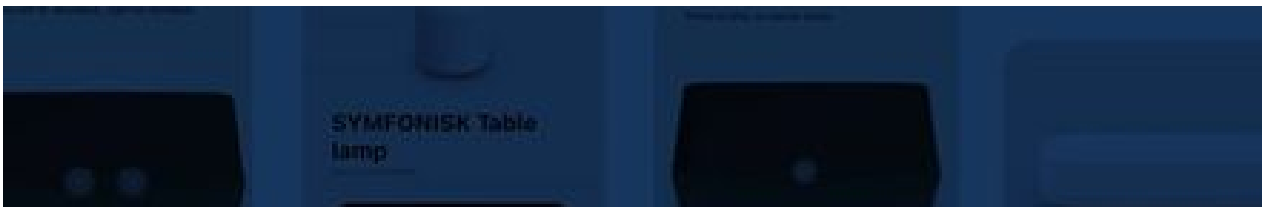
MACHINE LEARNING, OPEN SOURCE

Assembly still matters: Cortex-A53 vs M1

Read More →

December 6, 2021

Continue reading in User Experience:



SOFTWARE, USER EXPERIENCE

Renovating Setup, With Flutter

Read More →

May 4, 2022